



NETWORK UTILITY AWARE TRAFFIC LOADING BALANCING IN
BACKHAUL-CONSTRAINED CACHE-ENABLED SMALL CELL
NETWORKS WITH HYBRID POWER SUPPLIES

TAO HAN

NIRWAN ANSARI

TR-ANL-2014-007

SEP. 29, 2014

ADVANCED NETWORKING LABORATORY
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
NEW JERSEY INSTITUTE OF TECHNOLOGY

Network Utility Aware Traffic Loading Balancing in Backhaul-constrained Cache-enabled Small Cell Networks with Hybrid Power Supplies

Tao Han, *Student Member, IEEE*, and Nirwan Ansari, *Fellow, IEEE*

Advanced Networking Laboratory

Department of Electrical and Computer Engineering

New Jersey Institute of Technology, Newark, NJ, 07102, USA

Email: {th36, nirwan.ansari}@njit.edu

Abstract

Explosive data traffic growth leads to a continuous surge in capacity demands across mobile networks. In order to provision high network capacity, small cell base stations (SCBSs) are widely deployed. Owing to the close proximity to mobile users, SCBSs can effectively enhance the network capacity and offloading traffic load from macro BSs (MBSs). However, the cost-effective backhaul may not be readily available for SCBSs, thus leading to backhaul constraints in small cell networks (SCNs). Enabling cache in BSs may mitigate the backhaul constraints in SCNs. Moreover, the dense deployment of SCBSs may incur excessive energy consumption. To alleviate brown power consumption, renewable energy will be explored to power BSs. In such a network, it is challenging to dynamically balance traffic load among BSs to optimize the network utilities. In this paper, we investigate the traffic load balancing in backhaul-constrained cache-enabled small cell networks powered by hybrid energy sources. We have proposed a network utility aware (NUA) traffic load balancing scheme that optimizes user association to strike a tradeoff between the green power utilization and the traffic delivery latency. On balancing the traffic load, the proposed NUA traffic load balancing scheme considers the green power utilization, the traffic delivery latency in both BSs and their backhaul, and the cache hit ratio. The NUA traffic load balancing

scheme allows dynamically adjusting the tradeoff between the green power utilization and the traffic delivery latency. We have proved the convergence and the optimality of the proposed NUA traffic load balancing scheme. Through extensive simulations, we have compared performance of the NUA traffic load balancing scheme with other schemes and showed its advantages in backhaul-constrained cache-enabled small cell networks with hybrid power supplies.

I. INTRODUCTION

Owing to the proliferation of mobile devices and bandwidth greedy applications, mobile data traffic grows exponentially that has led to a continuous surge in network capacity demands [1]. Small cell base stations (SCBSs) are widely deployed to provision high network capacity [2]. SCBSs, with a small coverage area, can significantly improve the spectrum utilization in mobile networks and thus increase the network capacity [3]. However, owing to the disparate transmit powers and base station (BSs) capabilities, traditional traffic load balancing metrics such as the signal-to-interference-plus-noise ratio (SINR) and the received-signal-strength-indication (RSSI) may lead to a severe traffic load imbalance [2]. Hence, in order to fully exploit the capacity potential of small cell networks (SCNs), the traffic load balancing scheme should be well designed.

In mobile networks, traffic load balancing is achieved by executing user association process in which mobile users are assigned to base stations (BSs) for services. Various user association algorithms have been proposed to optimize the traffic load among BSs [2], [4]–[7]. Most of the existing solutions optimize the traffic load balancing in a mobile network with the implication that the air interface between BSs and mobile users is the bottleneck of the network. This implication is generally correct for BSs whose deployments are well planned. However, considering the potentially dense deployment of SCBSs, various backhaul solutions, e.g., xDSL, non-line-of-sight (NLOS) microwave, wireless mesh networks, rather than ideal backhaul such as optical fiber and LOS microwave, may be adopted [3]. As a result, backhaul instead of BSs may become the bottleneck of SCNs. To alleviate the backhaul constraints, content caching techniques have been explored to enable caching popular contents in BSs to reduce the traffic load in backhaul [8]–[11]. Therefore, it is desired to optimize the user association with the consideration of backhaul constraints and the performance of BSs' content cache system in SCNs.

Enhancing energy efficiency is also a critical task for next generation mobile networks [12], [13]. Although SCBSs consume less power than macro BSs (MBSs), the number of SCBSs will be orders of magnitude larger than that of MBSs for a wide scale network deployment. Hence, the overall power consumption of SCNs will be phenomenal. As energy harvesting technologies advance, renewable energy

such as sustainable biofuels, solar and wind energy can be utilized to power BSs [14]. Telecommunication companies such as Ericsson and Nokia Siemens have designed renewable energy powered BSs for mobile networks [15]. Define the electricity pulled from renewable energy systems and the power grid as green power and brown power, respectively. By adopting renewable energy powered BSs, mobile networks may further reduce their brown power consumption. However, since the electricity generated from renewable energy is not stable, green power may not be a reliable energy source for mobile networks. Therefore, future SCNs are likely to adopt hybrid energy supplies: brown power and green power. Green power is utilized to reduce the brown power consumption while brown power is utilized as a backup power source [16]. In order to optimize green power utilization, it is desirable to balance the traffic load according to the availability of green power. For instance, mobile networks may enable BSs with sufficient green power to serve more traffic load while reducing the traffic load of BSs consuming brown power [17]. Such traffic load balancing strategies, however, may not maximize network utilities such as the network capacity and the traffic delivery latency. Hence, a trade-off between the green power utilization and network utilities should be carefully evaluated in balancing traffic load among BSs.

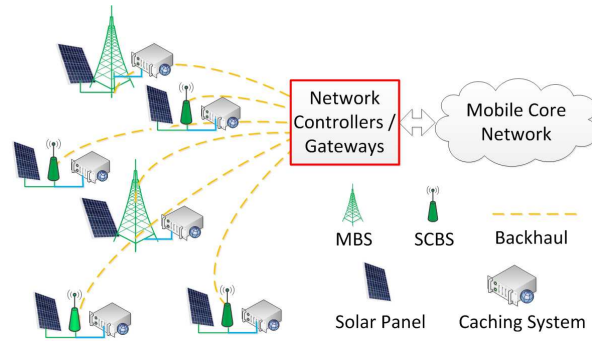


Fig. 1: The small cell network.

In this paper, we investigate the traffic load balancing in backhaul-constrained cache-enabled small cell networks with hybrid power supplies. The network architecture is shown in Fig. 1. The traffic load balancing in such a network requires the consideration of the green power utilization, BS capacity, backhaul constraints and the performance of cache systems. Therefore, we introduce four network utilities on balancing traffic load among BSs: 1) the green power utilization, 2) the traffic delivery latency in BSs, 3) the traffic delivery latency in backhaul, and 4) the cache hit ratio. The last three network utilities jointly determine the traffic delivery latency of the network. Thus, the awareness of these network utilities helps reduce the traffic delivery latency in the network. Since the green power utilization and the traffic

delivery latency are not optimized simultaneously in most scenarios, the tradeoff between the green power utilization and the traffic delivery latency should be determined based on network conditions. We propose the network utility aware (NUA) traffic load balancing scheme to adapt the user association according to the dynamics of these network utilities and strike an adjustable tradeoff between the brown power consumption and the traffic delivery latency. We prove the convergence and the optimality of the proposed NUA traffic load balancing scheme and validate its performance through extensive simulations.

The rest of the paper is organized as follows. In Section III, we define the system model and formulate the traffic load balancing problem. Section IV presents the proposed NUA traffic load balancing scheme and analyzes its properties. Section V shows the simulation results, and concluding remarks are presented in Section VI.

II. RELATED WORKS

Balancing traffic load in mobile networks has been extensively studied in recent years [2], [18]. In this section, we provide a briefly overview on existing traffic load balancing schemes. The most practical traffic load balancing approach is the cell range expansion (CRE) technique that biases users' receiving signal-to-interference-and-noise-ratios (SINRs) or data rates from some BSs to prioritize these BSs in associating with users [19]. Owing to the transmit power difference between MBSs and SCBSs, a large bias is usually given to SCBSs to offload users to small cells [2]. By applying CRE, a user associates with the BS from which the user receives the maximum biased SINR or biased data rate [4]. Deriving the optimal bias for BSs is challenging. Singh *et al.* [20] investigated the impact of the bias on network performances and provided a comprehensive analysis on traffic load balancing using CRE in heterogeneous mobile networks.

Optimization theory and game theory have been adopted to solve the traffic load balancing problem. Ye *et al.* [5] modeled the traffic load balancing problem as a utility maximization problem and developed distributed user association algorithms using the primal-dual decomposition. Kim *et al.* [6] proposed an α -optimal user association algorithm to achieve flow level load balancing under spatially heterogeneous traffic distribution. The proposed algorithm is based on convex optimization theory and may maximize different network utilities by selecting the value of α . Aryafar *et al.* [7] applied game theory to solve the traffic load balancing problem. The authors modeled the problem as a congestion game in which users are the players and user association decisions are the actions. Pantisano *et al.* [21] formulated the traffic load balancing problem in backhaul constrained SCNs as a one-to-many matching game between SCBSs and users and proposed a distributed algorithm based on the deferred acceptance scheme to obtain a

stable matching for mobile users.

Recognizing the green power utilization as one of the performance metrics when balancing the traffic load, Zhou *et al.* [22] proposed a handover parameter tuning algorithm for target cell selection, and a power control algorithm for coverage optimization to guide mobile users to access the BSs with renewable energy supply. Han and Ansari [17] proposed to optimize the utilization of green power for cellular networks by optimizing BSs' transmit powers. The proposed algorithm achieves significant brown power savings by scheduling the green power consumption along the time domain for individual BSs, and balancing the green power consumption among BSs. The authors have also proposed a user association framework that jointly optimizes the traffic delivery latency and the green power utilization [16], [23].

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we present the system model and the problem formulation. The system model includes the traffic and QoS model, and the energy model.

A. Traffic and QoS model

Denote \mathcal{B} as the set of BSs including both MBSs and SCBSs and \mathcal{A} as the coverage area of all BSs. Here, a BS refers to either a MBS or a SCBS. Since BSs are equipped with cache, users' data requests can be fulfilled by the cache system if the requested content is cached; otherwise, the requested content is retrieved from Internet. Retrieving contents from Internet generates traffic loads in a BS's backhaul. Therefore, we model the traffic delivery process as a queuing system as shown in Fig. 2.

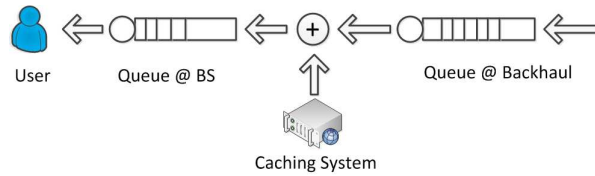


Fig. 2: The traffic delivery process as a queuing system.

The performance of a cache system is commonly evaluated based on the cache hit ratio that is defined as the ratio between the number of cache hits and the total requests observed over a period of time [24]. Many analytical models have been proposed to estimate the hit ratio of a cache system and various content caching strategies have been designed to optimize the performance of cache systems [24]–[28]. Thus, in the paper, we assume that the hit ratio of a cache system in a BS can be estimated for the time

duration of one user association process and denote $0 \leq \alpha_j \leq 1$ as the hit ratio of the cache system in BS j . Note that how to optimize and estimate the hit ratio is out of the scope of this paper.

Let a mobile user at location x associate with BS j . We assume that the traffic arrives at BS j 's backhaul toward the user according to a Poisson process with the arrival rate equaling to $\tilde{\lambda}(x)$, and the traffic loads per arrival (packet sizes per arrival) have an exponential distribution with the average traffic load of $\nu(x)$. We assume that the users associated with BS j are uniformly distributed in its coverage area and the traffic arrival processes are independent. For presentation simplicity, we further assume that no users share the same locations, i.e., only one user at location x . Since the traffic arrival toward a location is a Poisson process, the traffic arrival in BS j 's backhaul, which is the sum of the traffic arrivals from its coverage area, is also a Poisson process. Although BSs may adopt different access technologies as their backhaul, it is reasonable to assume the expected data rates of the backhaul are constant in the time duration of one user association process [21]. Since the traffic load per arrival follows an exponential distribution, the traffic delivery time (service time) of the backhaul is also an exponential distribution. Therefore, the traffic delivery in backhaul simply realizes an M/M/1 queuing system.

Denoting R_j as the average data rate of BS j 's backhaul. To fulfill the traffic demand of the user at location x , the required service time in BS j 's backhaul is

$$\tilde{\gamma}(x) = \frac{\nu(x)}{R_j}. \quad (1)$$

The average traffic load density generated by a user at location x in BS j 's backhaul is

$$\tilde{\varrho}_j(x) = \frac{\tilde{\lambda}(x)\nu(x)\eta_j(x)}{R_j} \quad (2)$$

Here, $\eta_j(x) = \{0, 1\}$ is an indicator function. If $\eta_j(x) = 1$, the user at location x is associated with BS j ; otherwise, the user is not associated with BS j . Since mobile users are uniformly distributed in the area, the traffic load in BS j 's backhaul can be expressed as

$$\tilde{\rho}_j = \int_{x \in \mathcal{A}} \tilde{\varrho}_j(x) dx. \quad (3)$$

According to the properties of the M/M/1 queue [29], the average waiting time for traffic load $\nu(x)$ in BS j 's backhaul is

$$\tilde{W}_j(x) = \frac{\tilde{\rho}_j \nu(x)}{R_j(1 - \tilde{\rho}_j)}. \quad (4)$$

Denote $\tilde{\mu}_j(x)$ as the latency ratio that measures how much time a user at location x must be sacrificed in waiting for per unit service time in BS j 's backhaul.

$$\tilde{\mu}_j(x) = \frac{\tilde{W}_j(x)}{\tilde{\gamma}(x)} = \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j}. \quad (5)$$

Since $\tilde{\mu}_j(x)$ only depends on the traffic load in BS j 's backhaul, all the users associated with BS j have the same latency ratio. Thus, we define

$$\tilde{\mu}_j(\tilde{\rho}_j) = \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j} \quad (6)$$

as the latency ratio of BS j 's backhaul. A smaller $\tilde{\mu}_j(\tilde{\rho}_j)$ indicates that BS j 's backhaul introduces less latency to its associated users.

According to Burke's Theorem [29], the traffic departure process at a BS's backhaul is a Poisson process with average departure rate equaling to the average traffic arrival rate. Therefore, the average traffic arrival rate in BS j toward a user at location x equals to $\tilde{\lambda}(x)$. Since the hit ratio of BS j 's cache system is α_j , the data traffic from the backhaul accounts for $(1 - \alpha_j)$ of the total traffic load toward the user at location x . Therefore, the average traffic arrival rate in BS j toward the user at location x is

$$\lambda(x) = \frac{\tilde{\lambda}(x)}{(1 - \alpha_j)}. \quad (7)$$

Since α_j is assumed to be a constant during one user association process, the traffic arrival process toward the user at location x is a Poisson process. In BS j , users at different locations may have different data rates depending on channel conditions. When associating with BS j , the user's data rate, $r_j(x)$, can be generally expressed as a logarithmic function of the perceived SINR, $SINR_j(x)$, according to the Shannon-Hartley Theorem [6],

$$r_j(x) = \log_2(1 + SINR_j(x)). \quad (8)$$

Here,

$$SINR_j(x) = \frac{P_j g_j(x)}{\sigma^2 + \sum_{k \in \mathcal{B}, k \neq j} P_k g_k(x)}. \quad (9)$$

Here, P_j is the transmission power of BS j , and σ^2 denotes the noise power level. Since the users' data rate is generally distributed, the service time in BS j follows a general distribution. Therefore, a BS's downlink transmission process realizes a M/G/1 processor sharing (PS) queue, in which multiple users share the BS's downlink radio resource [29].

In mobile networks, various downlink scheduling algorithms have been proposed to enable proper sharing of the limited radio resource in a BS. According to the scheduling algorithm, users may be assigned different priorities on sharing the radio resource. For analytical simplicity, we assume that mobile users are served based on the round robin (RR) fashion. Then, the average traffic load density at location x in BS j is calculated as

$$\varrho_j(x) = \frac{\lambda(x)\nu(x)\eta_j(x)}{r_j(x)} \quad (10)$$

The traffic load in BS j can be expressed as

$$\rho_j = \int_{x \in \mathcal{A}} \varrho_j(x) dx. \quad (11)$$

This value of ρ_j indicates the fraction of time during which BS j is busy. To fulfill the traffic demand of a user located at x , the required service time in BS j is

$$\gamma(x) = \frac{\nu(x)}{r_j(x)}. \quad (12)$$

Since the traffic delivery process in a BS realizes a M/G/1-RR queue, the average traffic delivery time for the user in BS j [29] is

$$T_j(x) = \frac{\nu(x)}{r_j(x)(1 - \rho_j)}. \quad (13)$$

The average waiting time for traffic load $\nu(x)$ in BS j is

$$W_j(x) = T_j(x) - \gamma(x) = \frac{\rho_j \nu(x)}{r_j(x)(1 - \rho_j)}. \quad (14)$$

Denote $\mu_j(x)$ as the latency ratio of BS j for a user at location x .

$$\mu_j(x) = \frac{W_j(x)}{\gamma(x)} = \frac{\rho_j}{1 - \rho_j}. \quad (15)$$

$\mu_j(x)$ only depends on the traffic load in BS j . Therefore, all the users associated with BS j have the same latency ratio. Thus, we define

$$\mu_j(\rho_j) = \frac{\rho_j}{1 - \rho_j} \quad (16)$$

as the latency ratio of BS j . A smaller $\mu_j(\rho_j)$ indicates that BS j introduces less latency to its associated users. In this paper, we use μ_j and $\tilde{\mu}_j$ to reflect the traffic delivery latency in BS j and its backhaul, respectively, we adopt $\mu_j(\rho_j) + \tilde{\mu}_j(\tilde{\rho}_j)$ as the QoS model that indicates the latency of delivering traffic through BS j .

B. Energy model

In the network, BSs have their own renewable energy systems (solar panels) for generating electricity. Meanwhile, BSs are connected with the power grid for electricity supplies. Thus, BSs are powered by hybrid energy sources: green power and brown power. A BS consumes brown power when green power is not sufficient. We assume that the green power systems in MBSs have a higher energy generation capacity than that of SCBSs because MBSs usually consume more energy than SCBSs owing to a relatively large coverage area. Fig. 3 shows a reference design of a hybrid energy powered BS [14].

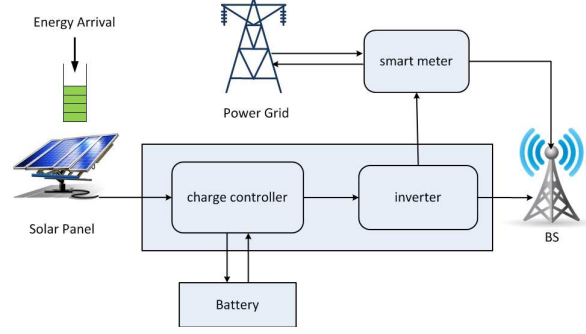


Fig. 3: A hybrid energy powered BS.

The charge controller optimizes the green power utilization based on the solar power intensity, the power consumption of BSs, and prices of energy drawn from the power grid. Based on the optimization, the charge controller determines how much green power should be utilized to power a BS during a specific time period, e.g., the time duration between two consecutive traffic load balancing procedures [16]. In this paper, we focus on how to balance traffic load among BSs to reduce the traffic delivery latency as well as the brown power consumption within the duration of a traffic balancing procedure. Investigating how to optimize the green power utilization over the time horizon is out of the scope of this paper. Thus, we assume that the amount of available green power for powering a BS within the duration of one user association process is given by the charge controller as a constant [16]. This assumption is reasonable because the traffic load balancing process is at a time scale of several minutes [6] while solar power generation is usually modeled at a time scale of a hour [30]. Denote e_j as the amount of green power for powering BS j in a user association procedure. If BS j 's power consumption is larger than e_j , the BS consumes brown power. Otherwise, we simply model the BS's brown power consumption are zero¹.

The BS's power consumption includes two parts: the static power consumption and the dynamic power consumption [31]. The static power consumption is the power consumption of a BS without carrying any traffic load. The dynamic power consumption refers to the additional power consumption incurred by traffic load in the BS, which can be well approximated by a linear function of the traffic load [31]. Denote p_j^s as BS j 's static power consumption. Then, BS j 's power consumption is

$$p_j = \beta_j \rho_j + p_j^s. \quad (17)$$

¹We do not consider the redistribution of the residual green power in our model, which is out of the scope of this paper.

Here, β_j is the load-power coefficient that reflects the relationship between BS j 's traffic load and its dynamic power consumption. The BS power consumption model can be adjusted to model the power consumption of either MBSs or SCBSs by incorporating and tweaking the static power consumption and the load-power coefficient [16]. The BS j 's brown power consumption is

$$p_j^b = \max(p_j - e_j, 0). \quad (18)$$

C. Problem formulation

In determining the user association, the network aims to not only enhance the network QoS by reducing the traffic delivery latency but also reduce the brown power consumption by improving the green power utilization. Owing to the dynamics of the data traffic and green power, the user association that minimizes the traffic delivery latency does not necessarily maximize the green power utilization. Thus, the traffic load balancing problem strives for a trade-off between the traffic delivery latency and the brown power consumption.

According to E.q. (18), brown power is consumed only when green power is not sufficient ($e_j < p_j$). Given e_j , the maximum traffic load can be supported by green power in BS j is

$$\hat{\rho}_j = \max(0, \min(\frac{e_j - p_j^s}{\beta_j}, 1 - \epsilon)). \quad (19)$$

Here, ϵ is an arbitrary small positive constant to guarantee $0 \leq \hat{\rho}_j < 1$. Define $\hat{\rho}_j$ as BS j 's green traffic capacity. When BS j 's traffic load is larger than $\hat{\rho}_j$, BS j consumes brown power. In this case, it is desirable to offload data traffic from BS j to alleviate its brown power consumption. Let

$$\bar{\rho}_j = \rho_j - \hat{\rho}_j. \quad (20)$$

When $\bar{\rho}_j > 0$, BS j 's traffic load is larger than its green power capacity and thus it is desirable to offload traffic from BS j to save brown power; when $\bar{\rho}_j < 0$, it is desirable to let BS j carry additional traffic load to enhance the usage of green power and thus reduce other BSs' brown power consumption. However, balancing traffic purely based on the energy consumption may lead to heavy traffic congestion that increases the traffic delivery latency in BSs. In order to strive a balance, we introduce latency weights for individual BSs. Denote

$$w_j(\rho_j) = e^{\kappa \bar{\rho}_j} \quad (21)$$

as BS j 's latency weight. $\kappa \geq 0$ is a system parameter that adjusts the value of the latency weight.

Aiming to save brown power as well as to reduce the traffic delivery latency of the network, the traffic load balancing problem is formulated as

$$\begin{aligned}
& \min_{\boldsymbol{\eta}} \quad \sum_{j \in \mathcal{B}} w_j(\rho_j) (\mu_j(\rho_j) + \tilde{\mu}_j(\tilde{\rho}_j)) \quad (22) \\
& \text{subject to :} \quad \rho_j = \int_{x \in \mathcal{A}} \frac{\lambda(x) \nu(x) \eta_j(x)}{r_j(x)} dx, \\
& \quad \quad \quad \tilde{\rho}_j = \int_{x \in \mathcal{A}} \frac{\tilde{\lambda}(x) \nu(x) \eta_j(x)}{R_j} dx, \\
& \quad \quad \quad 0 \leq \rho_j \leq 1 - \epsilon, \\
& \quad \quad \quad 0 \leq \tilde{\rho}_j \leq 1 - \epsilon, \\
& \quad \quad \quad \eta_j(x) = \{0, 1\}, \forall j \in \mathcal{B}, x \in \mathcal{A}. \quad (23)
\end{aligned}$$

Here, $\boldsymbol{\eta} = \{\eta_j | j \in \mathcal{B}\}$ and $\eta_j = \{\eta_j(x) | x \in \mathcal{A}\}$. Based on the formulation, if BS j has sufficient green power ($\tilde{\rho}_j \geq \rho_j$), $0 < w_j(\rho_j) \leq 1$; otherwise, $w_j(\rho_j) > 1$. A large latency weight grants a BS a high priority in minimizing Eq. (22) as compared with those of the BSs having a small latency weight. In other words, a large latency weight grants a BS a high priority in offloading traffic. As compared with $w_j(\rho_j) \leq 1$, $w_j(\rho_j) > 1$ enables BS j to achieve a smaller latency ratio. Since $\frac{d\mu_j(\rho_j)}{d\rho_j} > 0$ and $\frac{d\tilde{\mu}_j(\tilde{\rho}_j)}{d\tilde{\rho}_j} > 0$, a small latency ratio indicates that BS j carries a lighter traffic load, which is desirable for a BS which is consuming brown power ($w_j(\rho_j) \leq 1$). κ is a system parameter that enables the network dynamically controlling the trade-off between the brown power consumption and the traffic delivery latency.

IV. NETWORK UTILITY AWARE TRAFFIC LOAD BALANCING

In this section, we propose the network utility aware (NUA) traffic load balancing scheme and prove its properties. The network utilities considered in the traffic load balancing consist of 1) the green power utilization (brown power consumption), 2) the traffic delivery latency in BSs' backhaul, 3) the traffic delivery latency in BSs, and 4) the hit ratio of BSs' cache systems. The proposed network utility aware traffic load balancing scheme is able to adapt the traffic load among BSs and their backhauls according to the dynamics of these network utilities.

A. Traffic load balancing procedures

The traffic load balancing procedures can be implemented in either a distribute or a centralized fashion [16]. For a distributed traffic load balancing, users select their serving BSs based on the operating

parameters, e.g., traffic loads and data rates received from BSs. This will incur several interactions between BSs and users for updating the operating parameters and BS selections, respectively. For a centralized traffic load balancing scheme, the network collects the operating status information from both BSs and users, and determines the user association for individual users. A distributed traffic load balancing scheme can be implemented in a centralized fashion by leveraging virtualization techniques [16]. Thus, we propose a distributed traffic offloading procedure that includes four phases. The first phase is the initial user association and network utility measurements. When entering the network, a user simply attaches to any BS to retrieve network utility information. According to the initial user association, BSs measure their traffic load and estimate the traffic load in their backhaul. Based on the measurements, BSs, in the second phase, advertise their network utility information. Denote $\psi(\boldsymbol{\eta}) = \sum_{j \in \mathcal{B}} w_j(\rho_j)(\mu_j(\rho_j) + \tilde{\mu}_j(\tilde{\rho}_j))$. In the third phase, the users select their serving BSs according to the advertised network utility information and the downlink data rates to minimize $\psi(\boldsymbol{\eta})$. In the fourth phase, the BSs and users iteratively update their network utilities (the second phase) and BS selections (the third phase), respectively, until the user association converges.

B. The network utility aware user association

The network utility aware user association scheme consists of a user side algorithm and a BS side algorithm. The user side algorithm based on the network utility advertisement selects the optimal serving BS for individual users while the BS side algorithm updates individual BSs' network utility advertisements based on the user association. In designing the network utility aware user association scheme, we make the following assumptions:

- 1) We assume that the time scale of the traffic arrival and departure process is faster relative to that of BSs in advertising their network utility information. That is to say, BSs advertise their network utility information after the system exhibits the stationary performance.
- 2) We assume that the green power generation rate is consistent during the time period of establishing a stable user association [23].
- 3) We assume that all the BSs are synchronized and advertise their network utility simultaneously and the system parameter κ does not change during one user association process.
- 4) We assume that a BS's cache hit ratio is constant within the duration of one user association.

The feasible set for the traffic load balancing problem in Eq. (22) is

$$\begin{aligned}\mathcal{F} = \{\boldsymbol{\eta} | & 0 \leq \tilde{\rho}_j \leq 1 - \epsilon, \\ & 0 \leq \rho_j \leq 1 - \epsilon, \sum_{j \in \mathcal{B}} \eta_j(x) = 1, \\ & \eta_j(x) = \{0, 1\}, \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}\end{aligned}\quad (24)$$

Since $\eta_j(x) = \{0, 1\}$, $\psi(\boldsymbol{\eta})$ is not continuous differentiable. In order to derive the user side algorithm and the BS side algorithm for the NUA traffic load balancing scheme, we relax the feasible set by letting $0 \leq \eta_j(x) \leq 1$. Then, the relaxed feasible set is

$$\begin{aligned}\tilde{\mathcal{F}} = \{\boldsymbol{\eta} | & 0 \leq \tilde{\rho}_j \leq 1 - \epsilon, \\ & 0 \leq \rho_j \leq 1 - \epsilon, \sum_{j \in \mathcal{B}} \eta_j(x) = 1, \\ & 0 \leq \eta_j(x) \leq 1, \forall j \in \mathcal{B}, \forall x \in \mathcal{A}\}\end{aligned}\quad (25)$$

After presenting the NUA traffic load balancing scheme, we will prove that the proposed scheme achieves an optimal user association in the feasible set of the traffic load balancing problem.

We define the time interval between two consecutive network utility advertisements as a time slot. Let $\eta_j^k(x)$ denote whether a user at location x associates with BS j in the k th time slot. Denote $\rho_j(k)$ and $\tilde{\rho}_j(k)$ as the traffic load in BS j and its backhaul in the k th time slot, respectively. Let

$$\begin{aligned}\phi_j(k) &= \frac{d\psi(\boldsymbol{\eta})}{d\eta_j(x)} \\ &= \frac{\lambda(x)\nu(x)}{r_j(x)} e^{\kappa(\rho_j(k) - \hat{\rho}_j)} \left(\frac{\kappa\rho_j(k)}{1 - \rho_j(k)} + \frac{\kappa\tilde{\rho}_j(k)}{1 - \tilde{\rho}_j(k)} + \frac{1}{(1 - \rho_j(k))^2} \right) \\ &\quad + \lambda(x)\nu(x) e^{\kappa(\rho_j(k) - \hat{\rho}_j)} \frac{1 - \alpha_j}{R_j(1 - \tilde{\rho}_j(x))^2}.\end{aligned}\quad (26)$$

Define

$$\theta_j^a(k) = e^{\kappa(\rho_j(k) - \hat{\rho}_j)} \left(\frac{\kappa\rho_j(k)}{1 - \rho_j(k)} + \frac{\kappa\tilde{\rho}_j(k)}{1 - \tilde{\rho}_j(k)} + \frac{1}{(1 - \rho_j(k))^2} \right) \quad (27)$$

and

$$\theta_j^b(k) = e^{\kappa(\rho_j(k) - \hat{\rho}_j)} \frac{1 - \alpha_j}{R_j(1 - \tilde{\rho}_j(x))^2}. \quad (28)$$

Since $\theta_j^a(k)$ and $\theta_j^b(k)$ are calculated based on BS j 's network utility, we define $\theta_j^a(k)$ and $\theta_j^b(k)$ as the network utility information advertised by BS j in the k th time slot.

1) *The User Side Algorithm:* At the beginning of the k th time slot, BSs broadcast their network utility advertisements, e.g., $\theta_j^a(k)$ and $\theta_j^b(k)$, to users. The BS selection rule for a user at location x is

$$b^k(x) = \arg \max_{j \in \mathcal{B}} \frac{r_j(x)}{\theta_j^a(k) + r_j(x)\theta_j^b(k)}. \quad (29)$$

Here, $b^k(x)$ is the index of the BS selected by the user. Therefore,

$$\eta_j^k(x) = \begin{cases} 1, & \text{for } j = b^k(x), \forall x \in \mathcal{A} \\ 0, & \text{for } j \neq b^k(x), \forall x \in \mathcal{A}, \end{cases} \quad (30)$$

2) *The BS Side Algorithm:* After mobile users select their associating BSs, the user association in BS j , η_j^k , is updated. Give the user association, BS j updates its network utility advertisement. BS j calculates an intermediate user association $\bar{\eta}_j^k = \{\bar{\eta}_j^k | j \in \mathcal{B}\}$ as

$$\bar{\eta}_j^k = (1 - \delta^k)\eta_j^k + \delta^k\bar{\eta}_j^{k-1}. \quad (31)$$

Here, $0 < \delta^k < 1$ is an exponential averaging parameter. With the intermediate user association, BS j calculates the intermediate traffic load in the BS and its backhaul. The intermediate traffic load in BS j 's backhaul is

$$\tilde{\rho}_j(k+1) = \int_{x \in \mathcal{A}} \frac{\lambda(x)\nu(x)\bar{\eta}_j^k(x)}{R_j} dx, \quad (32)$$

and intermediate traffic load in BS j is

$$\rho_j(k+1) = \int_{x \in \mathcal{A}} \frac{\lambda(x)\nu(x)\bar{\eta}_j^k(x)}{r_j(x)} dx. \quad (33)$$

Based on the intermediate traffic load in both the BS and its backhaul, BS j calculates its network utility advertisements, $\theta_j^a(k+1)$ and $\theta_j^b(k+1)$, using Eqs. (27) and (28).

C. The properties of the network utility aware user association

In this subsection, we prove the convergence and the optimality of the proposed NUA traffic load balancing scheme. Since users select serving BSs based on BSs' network utility advertisements, the user association converges when BSs' network utility advertisements are stabilized. A BS's network utility advertisements are determined by its intermediate traffic loads in the BS and its backhaul. On calculating the intermediate traffic loads, the intermediate user association is the only variable. Therefore, when the intermediate user association converges, the intermediate traffic load is stabilized and so are the network utility advertisements. Therefore, we first prove any BS's network utility advertisements are stabilized by proving that its intermediate user association converges and then show that the user association based on the stabilized network utility advertisements minimizes $\psi(\boldsymbol{\eta})$.

Lemma 1. *The relaxed feasible set $\tilde{\mathcal{F}}$ is a convex set.*

Proof: The lemma is proved by showing that the set $\tilde{\mathcal{F}}$ contains any convex combination of the user association vector η . ■

Lemma 2. *$\psi(\eta)$ is a convex function of η when η is defined in $\tilde{\mathcal{F}}$.*

Proof: The lemma can be proved by showing $\nabla^2\psi(\eta) > 0$ when η is defined in $\tilde{\mathcal{F}}$. ■

Let $\bar{\eta}^k = \{\bar{\eta}_j^k | j \in \mathcal{B}\}$ and $\Delta\bar{\eta}^k = \bar{\eta}^k - \bar{\eta}^{k-1}$.

Lemma 3. *When $\Delta\bar{\eta}^k \neq 0$, $\Delta\bar{\eta}^k$ provides a descent direction of $\psi(\bar{\eta})$ at $\bar{\eta}^k$.*

Proof: Since $0 \leq \bar{\eta}_j^k \leq 1, \forall k, \forall j \in \mathcal{B}$, $\bar{\eta}$ is defined in $\tilde{\mathcal{F}}$. According to Lemmas 1 and 2, $\psi(\bar{\eta})$ is a convex function of $\bar{\eta}$. Hence, the lemma can be proved by showing $\langle \nabla\psi(\bar{\eta})|_{\bar{\eta}=\bar{\eta}^k}, \Delta\bar{\eta}^k \rangle < 0$.

$$\langle \nabla\psi(\bar{\eta})|_{\bar{\eta}=\bar{\eta}^k}, \Delta\bar{\eta}^k \rangle \quad (34)$$

$$\begin{aligned} &= \int_{x \in \mathcal{A}} \sum_{j \in \mathcal{B}} \lambda(x) \nu(x) (\bar{\eta}_j^k(x) - \bar{\eta}_j^{k-1}(x)) \left(\frac{\theta_j^a(k)}{r_j(x)} + \theta_j^b(k) \right) \\ &= (1 - \delta^k) \int_{x \in \mathcal{A}} \lambda(x) \nu(x) \sum_{j \in \mathcal{B}} (\eta_j^k(x) - \bar{\eta}_j^{k-1}(x)) \frac{\theta_j^a(k) + r_j(x) \theta_j^b(k)}{r_j(x)} \end{aligned} \quad (35)$$

Since

$$\eta_j^k(x) = \begin{cases} 1, & \text{for } j = b^k(x) \\ 0, & \text{for } j \neq b^k(x), \end{cases} \quad (36)$$

$$\sum_{j \in \mathcal{B}} (\eta_j^k(x) - \bar{\eta}_j^{k-1}(x)) \frac{\theta_j^a(k) + r_j(x) \theta_j^b(k)}{r_j(x)} \leq 0. \quad (37)$$

Because $0 < \delta^k < 1$ and $\Delta\bar{\eta}_j^k \neq 0$,

$$\sum_{j \in \mathcal{B}} (\eta_j^k(x) - \bar{\eta}_j^{k-1}(x)) \frac{\theta_j^a(k) + r_j(x) \theta_j^b(k)}{r_j(x)} < 0. \quad (38)$$

Thus, $\langle \nabla\psi(\bar{\eta})|_{\bar{\eta}=\bar{\eta}^k}, \Delta\bar{\eta}^k \rangle < 0$. ■

Denote $\bar{\eta}^*$ as the optimal intermediate user association.

Lemma 4. *When $\bar{\eta}^k \neq \bar{\eta}^*$, $\bar{\eta}^k \in \tilde{\mathcal{F}}$, there exists $0 < \delta^k < 1$ such that $\psi(\bar{\eta}^k) < \psi(\bar{\eta}^{k-1})$.*

Proof: Since

$$\begin{aligned}\Delta \bar{\eta}^k &= \bar{\eta}^k - \bar{\eta}^{k-1} \\ &= (1 - \delta^k)(\eta^k - \bar{\eta}^{k-1}),\end{aligned}\tag{39}$$

$(\eta^k - \bar{\eta}^{k-1})$, according to Lemma 3, provides the descent direction for searching the optimal value in the iterations while $(1 - \delta^k)$ indicates the search step in the k th iteration. Since $\bar{\eta}^k \neq \bar{\eta}^*$, there exists $0 < \delta^k < 1$ that enables $\psi(\bar{\eta}^k) < \psi(\bar{\eta}^{k-1})$ ■

Theorem 1. *If the traffic load balancing problem is feasible² and δ^k is properly selected, $\bar{\eta}^k = (1 - \delta^k)\eta^k + \delta^k\bar{\eta}^{k-1}$ converges to $\bar{\eta}^*$.*

Proof: Since 1) $\bar{\eta}_j^k - \bar{\eta}_j^{k-1}$ is a descent direction of $\psi(\bar{\eta})$ at $\bar{\eta}^k$ and 2) δ^k is properly selected such that $\psi(\bar{\eta}^k) < \psi(\bar{\eta}^{k-1})$, the mapping, $\bar{\eta}^k = (1 - \delta^k)\eta^k + \delta^k\bar{\eta}^{k-1}$, keep decreasing $\psi(\bar{\eta})$. Since $\psi(\bar{\eta}) \geq 0$, $\bar{\eta}^k$ will eventually converge. According to Lemma 4, $\bar{\eta}^k$ converges to $\bar{\eta}^*$. Otherwise, $\psi(\bar{\eta})$ can be further reduced. ■

Corollary 1. *Any BS's network utility advertisements, $\theta_j^a(k)$ and $\theta_j^b(k)$, $j \in \mathcal{B}$, are stabilized.*

Proof: $\theta_j^a(k)$ and $\theta_j^b(k)$, $j \in \mathcal{B}$, are calculated by the traffic load in BS j and its backhaul, respectively. When the intermediate user association converges, the traffic loads are determined. As a result, individual BS's network utility advertisements are stabilized. ■

Theorem 2. *Given that the traffic load balancing problem is feasible, the user association, $\eta_j^* = \{\eta_j^*(x) | \eta_j^*(x) = \{0, 1\}, x \in \mathcal{A}\}$, $j \in \mathcal{B}$, based on the stabilized network utility advertisements is an optimal solution to the traffic load balancing problem.*

Proof: Denote $\theta_j^a(*)$ and $\theta_j^b(*)$ as BS j 's stabilized network utility advertisements. Let $\eta^* = \{\eta_j^* | j \in \mathcal{B}\}$ and $\eta = \{\eta_j | j \in \mathcal{B}\}$. Here, $\eta_j = \{\eta_j(x) | \eta_j(x) = \{0, 1\}, x \in \mathcal{A}\}$. Suppose η to be arbitrary user association in the feasible set \mathcal{F} that is not equal to η^* .

²The problem is feasible when the feasible set of the problem is not empty.

$$\begin{aligned}
& \langle \nabla \psi(\boldsymbol{\eta}) |_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}, \boldsymbol{\eta} - \boldsymbol{\eta}^* \rangle \\
&= \int_{x \in \mathcal{A}} \sum_{j \in \mathcal{B}} \lambda(x) \nu(x) (\eta_j(x) - \eta_j^*(x)) \left(\frac{\theta_j^a(*)}{r_j(x)} + \theta_j^b(*) \right) \\
&= \int_{x \in \mathcal{A}} \lambda(x) \nu(x) \sum_{j \in \mathcal{B}} (\eta_j(x) - \eta_j^*(x)) \left(\frac{\theta_j^a(*)}{r_j(x)} + \theta_j^b(*) \right)
\end{aligned} \tag{40}$$

Since

$$b^*(x) = \arg \max_{j \in \mathcal{B}} \frac{r_j(x)}{\theta_j^a(*) + r_j(x) \theta_j^b(*)} \tag{41}$$

and

$$\eta_j^*(x) = \begin{cases} 1, & \text{for } j = b^*(x) \\ 0, & \text{for } j \neq b^*(x), \end{cases} \tag{42}$$

$$\sum_{j \in \mathcal{B}} \eta_j(x) \left(\frac{\theta_j^a(*)}{r_j(x)} + \theta_j^b(*) \right) \geq \sum_{j \in \mathcal{B}} \eta_j^*(x) \left(\frac{\theta_j^a(*)}{r_j(x)} + \theta_j^b(*) \right). \tag{43}$$

Hence, $\langle \nabla \psi(\boldsymbol{\eta}) |_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}, \boldsymbol{\eta} - \boldsymbol{\eta}^* \rangle \geq 0$. Therefore, $\boldsymbol{\eta}^*$ is an optimal solution to the UA problem. \blacksquare

D. The adaptation of the energy-latency tradeoff

The system parameter, κ , controls the tradeoff between the green power utilization and the traffic delivery latency. When $\kappa = 0$, $w_j(\rho_j) = 1$. In this case, the green power utilization is not modeled in the objective function. Thus, the NUA traffic load balancing scheme determines the user association based only on the traffic delivery latency. As κ increases, the awareness of green power utilization in determining the user association enhances. In other words, with a larger κ , the green power utilization plays a more important role in determining the user association. If κ is large enough, the NUA traffic load balancing scheme achieves a user association that approximates the user association that only cares about the green power utilization.

V. SIMULATION AND PERFORMANCE EVALUATION

A. Simulation setup

We set up system level simulations to investigate the performance of the NUA traffic load balancing scheme for the downlink traffic load balancing in backhaul constrained SCNs. In the simulation, three MBSs and seven SCBSs are randomly deployed in a $2000m \times 2000m$ area. The total bandwidth is 10 MHz and the frequency reuse factor is one. The channel propagation model is based on COST 231 Walfisch-Ikegami [32]. The channel model and parameters are summarized in Table I. Here, PL_{MBS} and

TABLE I: Channel Model and Parameters

Parameters	Value
PL_{MBS} (dB)	$PL_{MBS} = 128.1 + 37.6 \log_{10}(d)$
PL_{SCBS} (dB)	$PL_{SCBS} = 38 + 10 \log_{10}(d)$
Rayleigh fading	9 dB
Shadowing fading	5 dB
Antenna gain	15 dB
Noise power level	-174 dBm
Receiver sensitivity	-123 dBm

PL_{SCBS} are the path loss between the users and MBSs and SCBSs, respectively. d is the distance between users and BSs. The transmit power of an MBS and an SCBS are 43 dBm and 33 dBm, respectively.

TABLE II: The Average Cache Hit Ratio

MBS 1	MBS 2	MBS 3	SCBS 4	SCBS 5	SCBS 6	SCBS 7	SCBS 8	SCBS 9	SCBS 10
0.27	0.12	0.28	0.12	0.17	0.22	0.22	0.24	0.24	0.19

The static power consumptions of an MBS and an SCBS are 750 W and 37 W, respectively [31]. The load-power coefficients of the MBS and the SCBS are 500 and 4, respectively [31]. The solar cell power efficiency is 17.4% [33]. We assume that the weather condition is the standard condition which specifies a temperature of 25 °C, an irradiance of 1000 W/m², and an air mass of 1.5 spectrum [34]. Thus, the green power generation rate is 174 W/m². The solar panel sizes are randomly selected but ensure the green power generation capacity of MBSs from 750 w to 1300 w while that of SCBSs from 37 w to 48 w. BSs' energy-latency coefficients are set to be the same. In the simulation, the average data rate of SCBSs' backhaul is 5 Mbps. The hit ratio³ of BSs' cache system is shown in Table II.

B. Traffic load balancing algorithms and network utility awareness

In the simulations, we investigate the performance of traffic load balancing schemes with different levels of network utility awareness. The network utilities considered in this paper are green power, the traffic delivery latency in BSs, the traffic delivery latency in backhaul, and the cache hit ratio. We implement

³The cache hit ratio is randomly selected from 0.1 to 0.3. For the analytical simplicity, we fix the hit ratio of BSs in the simulations.

TABLE III: Network Utility Aware User Association Schemes

UA Scheme	green power	BS Latency	Backhaul Latency	Cache
vGALA ($\kappa = 0$)		x		
vGALA ($\kappa = 6, \theta = 1$)	x			
vGALA ($\kappa = 4, \theta = 0.5$)	x	x		
NUA ($\kappa = 0, \alpha_j = 0$)		x	x	
NUA ($\kappa = 0, \alpha_j > 0$)		x	x	x
NUA ($\kappa = 2, \alpha_j = 0$)	x	x	x	
NUA ($\kappa = 2, \alpha_j > 0$)	x	x	x	x
DRB-NU	x	x	x	x

three traffic load balancing schemes in the simulations. The first scheme is vGALA [16]. Adapting the parameters of vGALA (κ and θ), we realize the three traffic load balancing schemes: 1) BS latency aware, 2) green power aware, and 3) BS latency and green power aware. The second scheme is the NUA traffic load balancing scheme that simulates four traffic load balancing schemes: 1) BS latency and backhaul latency aware, 2) BS latency, backhaul latency and cache hit ratio aware, 3) BS latency, backhaul latency and green power aware, and 4) all network utilities aware. For the NUA scheme, when $\alpha_j = 0, \forall j \in \mathcal{B}$, a BS estimates the traffic delivery latency in backhaul purely based on the traffic arrival rates in the BS. In fact, if the cache system is considered, the traffic load in backhaul should be less than that in the BS. Therefore, the NUA scheme with $\alpha_j = 0, \forall j \in \mathcal{B}$, simulates the cache unaware traffic load balancing scheme. The third scheme is the data rate bias (DRB) scheme [19]. In the implementation, we assume that BSs in the same tier have the data rate bias. MBSs are in the first tier while SCBSs are in the second tier. In the data rate bias scheme, a user selects the serving BS to maximize the biased data rate. The data rate bias of an MBS is set to one. We vary the data rate bias of an SCBS to investigate the performance of the scheme. We set $\psi(\boldsymbol{\eta})$ as the performance metric for selecting the optimal data rate bias. Thus, the implemented DRB scheme is aware of all network utilities and is referred to as DRB-NU (the data rate bias with network utility awareness). The network utility awareness of the schemes with different settings are shown in Table III.

C. Simulation results

Fig. 4 shows the convergence of the NUA scheme and its energy-latency tradeoff with different κ . Fig. 4a shows the the value of $\psi(\boldsymbol{\eta})$ converges with less than 100 iterations, and so do the traffic delivery

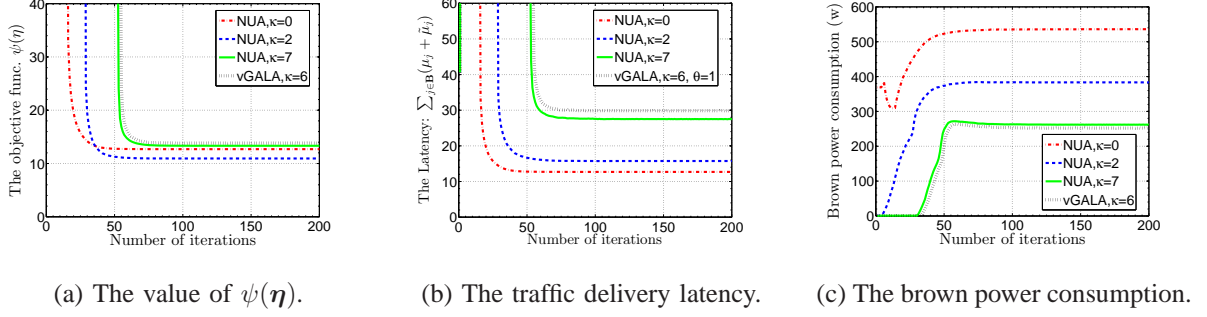


Fig. 4: The performance of the NUA scheme with different κ .

latency (Fig. 4b) and the brown power consumption (Fig. 4c). Figs. 4b and 4c show the energy-latency tradeoff. As κ increases, the network emphasizes the green power utilization in determining the user association. As a result, with a large κ , e.g., $\kappa = 7$, the network consumes less brown power at the cost of introducing additional traffic delivery latency. The vGALA with a large κ and θ , e.g., $\kappa = 6$ and $\theta = 1$, realizes the user association that is only aware of green power utilization [16]. Fig. 4c shows that as κ increases, the performance of the NUA scheme in terms of the brown power consumption approaches that of the traffic load balancing scheme that optimizes the green power utilization (only aware of green power utilization).

Fig. 5 shows the performance of the NUA scheme versus different solar panel efficiency. Fig. 5b shows that the brown power consumption reduces as the solar panel efficiency increases. This is because a higher solar panel efficiency enables solar panels to generate a larger amount of electricity and thus lessen the brown power consumption. As shown in Fig 5c, the performance of the traffic delivery latency divides into four regions. In first region (R1), the traffic delivery latency does not change. This is because the green power generated in individual BSs is less than their static power consumption when the solar panel efficiency is in R1. In other words, the green capacity of all BSs is zero when the solar panel efficiency is within R1. As a result, increasing the solar panel efficiency in R1 does not impact the traffic delivery latency, and neither does the value of $\psi(\eta)$. In the second region, as shown in Fig. 5c, the traffic delivery latency increases as the solar panel efficiency increases. When the solar panel efficiency is within this region, the network trades the traffic delivery latency for reducing the brown power consumption. In the third region (R3), the network trades the power consumption for reducing the traffic delivery latency. This can be seen from Fig. 5b. The rate of brown power consumption reduction decreases when the solar panel efficiency is about 17% (the start point of R3) as shown in Fig. 5c. This indicates that the network

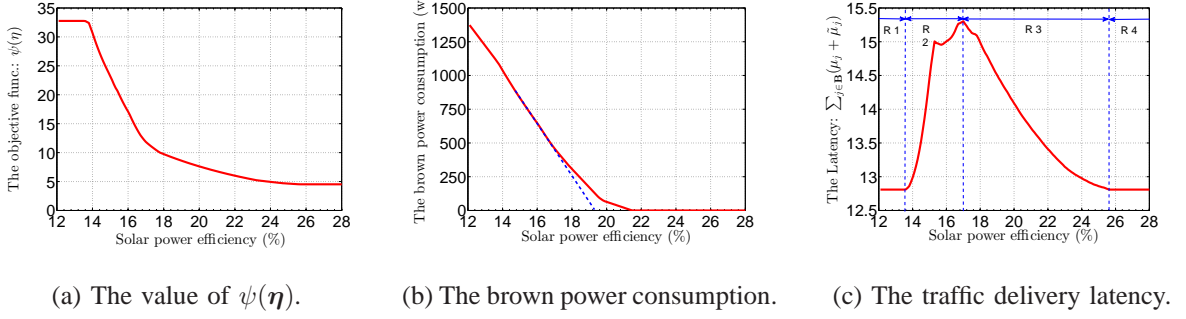


Fig. 5: The performance of the NUA scheme versus the solar panel efficiency.

emphasizes on reducing the traffic delivery latency when the solar panel efficiency is within R3. In both R2 and R3, the energy-latency tradeoff reduces the value of $\psi(\eta)$ as shown in Fig. 5a. When the solar panel efficiency is within the fourth region (R4), the solar panel efficiency is high enough to enable zero brown power consumption in all BSs while minimizing the traffic delivery latency.

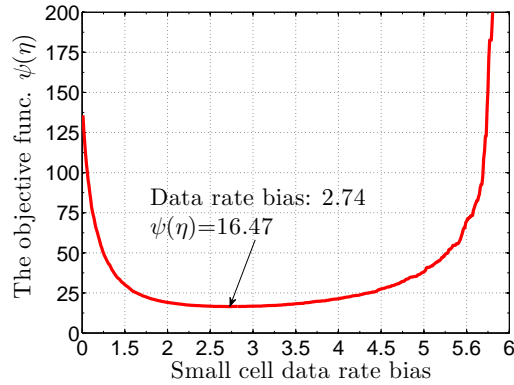


Fig. 6: The value of $\psi(\eta)$ versus data rate bias.

Fig. 6 shows the value of $\psi(\eta)$ versus the small cell data rate biases under the DRB-NU scheme. The value of $\psi(\eta)$ is minimized when the small cell data rate bias equals to 2.74. Given the data rate bias, the network's traffic delivery latency is 17.24 and the brown power consumption is 480.9 w. Under the NUA scheme with $\kappa = 2$, the network's traffic delivery latency and the brown power consumption are 15.74 and 383.35 w, respectively. Therefore, as compared with the DRB-NU scheme, the NUA scheme reduces the traffic delivery latency and the brown power consumption by 8.7% and 20.28%, respectively. The NUA scheme has achieved enhanced performance because it allows individual BSs to adapt their network utility advertisements while the DRB-NU scheme only allows to change the data rate bias for

an entire tier rather than for individual BSs. Another drawback of the DRB-NU scheme is that it does not dynamically respond to the network utility changes. The small cell data rate bias is optimized based on the previous instead of current network conditions, e.g., traffic intensities, backhaul constraints, and green power availabilities.

In Figs. 7, 9, 8, and 10, we compare the performance of three traffic load balancing schemes with varying network conditions. The first scheme is the green power and BS latency aware traffic load balancing scheme realized by vGALA with $\kappa = 4$ and $\theta = 0.5$ [16]. The second one is the NUA scheme that is aware of all network utilities. The third one, referred to as NUA-NC (no cache), is the NUA without awareness of the cache hit ratio. This scheme is realized by the NUA scheme with $\alpha_j = 0, \forall j \in \mathcal{B}$. Fig. 7 shows the performance of these traffic load balancing schemes versus different backhaul data rates. When the backhaul data rate is very low, e.g., less than 5 Mbps in the simulation, the value of $\psi(\boldsymbol{\eta})$ and the traffic delivery latency under vGALA is very large. This indicates that, without the awareness of backhaul data rates, the traffic load balancing under vGALA congests some BSs in the network. The brown power consumption under vGALA does not change versus the backhaul data rates because vGALA does not consider the traffic delivery latency in backhaul as a performance metric in determining the user association. As the backhaul data rates increase, the value of $\psi(\boldsymbol{\eta})$ under these schemes converges because the backhaul constraint is gradually mitigated. However, the NUA scheme achieves smaller traffic delivery latency as compared with the vGALA scheme because of the awareness of the traffic delivery latency in backhaul.

In the simulation, the value of $\psi(\boldsymbol{\eta})$ is minimized by the NUA scheme. However, the NUA-NC scheme achieves the minimal traffic delivery latency as shown in Fig. 7c. This is because the NUA scheme aims to minimize the value of $\psi(\boldsymbol{\eta})$ and thus strikes a tradeoff between the traffic delivery latency and the brown power consumption. As a result, compared with the NUA-NC scheme, the NUA scheme consumes less brown power at the cost of an increase of the traffic delivery latency.

The BSs' coverage areas under these schemes are shown in Fig. 8. The NUA-NC scheme is aware of the backhaul limitation and thus reduces the coverage area of the backhaul constrained SCBS, e.g., SCBS 8. However, owing to the unawareness of the cache hit ratio, the NUA-NC scheme overestimates the traffic load in backhaul and constrains the coverage area of SCBSs, e.g., SCBS 8. The NUA scheme, being aware of the cache hit ratio of individual BSs, accurately estimates the traffic load in the backhaul and thus derives optimal coverage areas for BSs, e.g., increasing the coverage area of SCBS 8, to minimize the value of $\psi(\boldsymbol{\eta})$.

As shown in Fig. 9, when the backhaul data rate of a SCBS changes, e.g., R_5 reduces from 5 Mbps

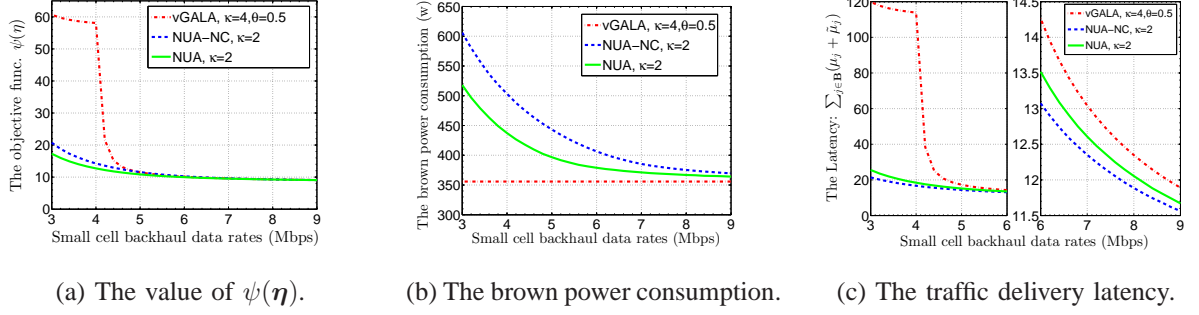


Fig. 7: The performance of the traffic load balancing schemes versus the backhaul data rates.

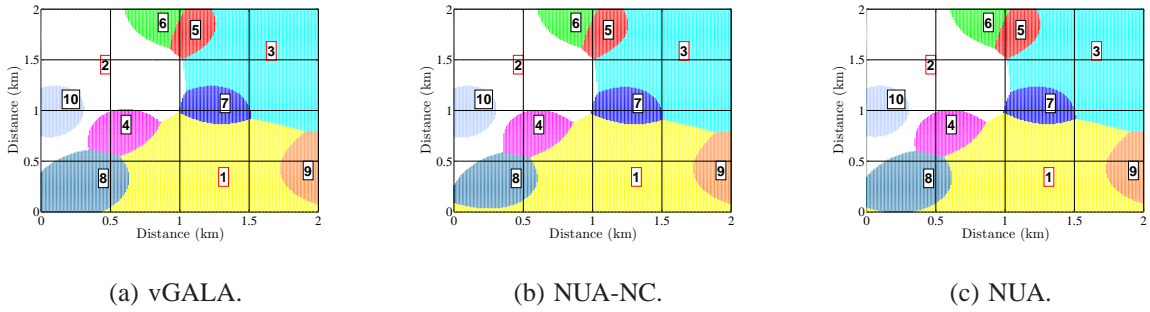


Fig. 8: The coverage areas of different schemes ($R_{4-10} = 5 \text{ Mbps}$).

to 1 Mbps, the NUA and NUA-NC scheme are able to adapt the traffic load balancing according to the backhaul data rate changes. However, the vGALA scheme, without the awareness of backhaul data rate, incurs excessive traffic delivery latency which is 667% of the traffic delivery latency of the NUA scheme as shown in Fig.9c. As shown in Fig. 10, both the NUA and NUA-NC schemes are able to reduce the coverage area of SCBS 5. The NUA-NC scheme, because of the unawareness of the cache hit ratio, shrinks the coverage area more than the NUA scheme does.

Fig. 11 shows the the impact of the cache awareness on the traffic delivery latency. In the simulation, we set $\kappa = 0$ for both the NUA scheme and the NUA-NC scheme to focus on the performance of the traffic delivery latency. Thus, both schemes are unaware of the green power utilization. As shown in Fig. Fig. 11, when the backhaul data rate is small, the cache awareness helps to reduce the traffic delivery latency.

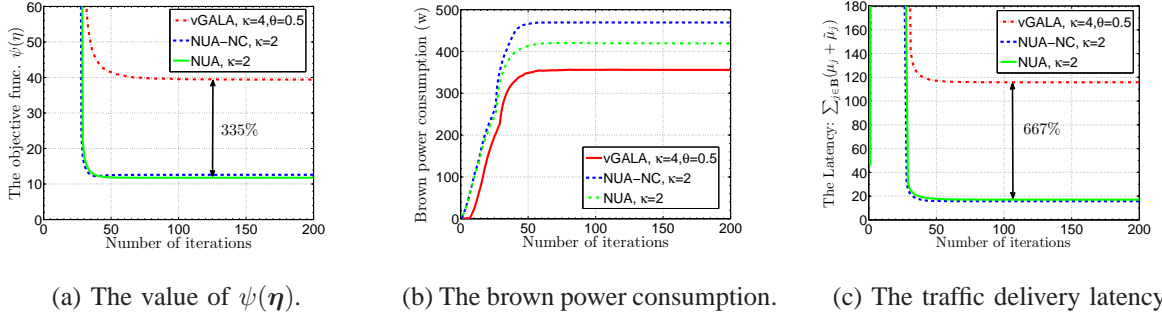


Fig. 9: The performance comparison ($R_5 = 1 \text{ Mbps}$ and $R_{4,6-10} = 5 \text{ Mbps}$).

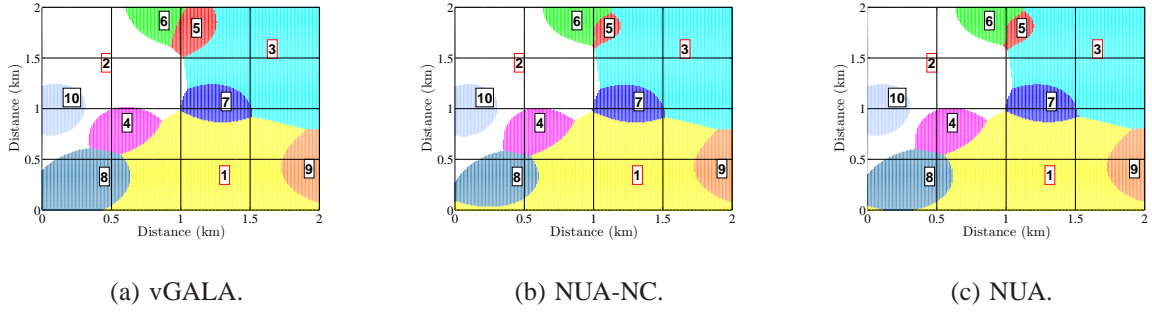


Fig. 10: The coverage areas of different schemes ($R_5 = 1 \text{ Mbps}$ and $R_{4,6-10} = 5 \text{ Mbps}$).

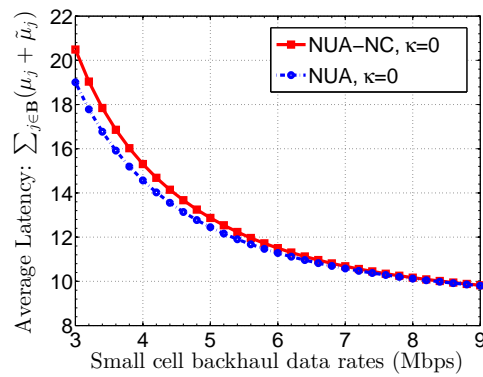


Fig. 11: The impact of the cache awareness on the traffic delivery latency.

VI. CONCLUSION

In this paper, we have proposed a network utility aware (NUA) traffic load balancing scheme for backhaul-constrained cache-enabled SCNs with hybrid power supplies. During the procedure of establishing user associations, the NUA traffic load balancing scheme considers four network utilities: green power utilization, the traffic delivery latency in BSs, the traffic delivery latency in backhaul, and the cache hit ratio. By optimizing the user association, the NUA traffic load balancing scheme strikes a tradeoff between the green power utilization and the traffic delivery latency in the network. The NUA traffic load balancing scheme adapts the user association according to the dynamics of green power, BS capacity, backhaul data rates, and the cache hit ratio. It significantly reduces the traffic delivery latency when the network is constrained by the backhaul data rate. Moreover, by adjusting the system parameters, e.g., κ , the NUA scheme is able to adjust the tradeoff between the brown power consumption and the traffic delivery latency.

REFERENCES

- [1] T. Han, N. Ansari, M. Wu, and H. Yu, "On accelerating content delivery in mobile networks," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 3, pp. 1314–1333, Third 2013.
- [2] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: old myths and open problems," *Wireless Communications, IEEE*, vol. 21, no. 2, pp. 18–25, April 2014.
- [3] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, H. Tang, X. Shen, N. Yang, and N. Li, "Trends in small cell enhancements in lte advanced," *Communications Magazine, IEEE*, vol. 51, no. 2, pp. 98–105, February 2013.
- [4] H.-S. Jo, Y. J. Sang, P. Xia, and J. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.
- [5] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 6, pp. 2706–2716, June 2013.
- [6] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α -optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, Feb. 2012.
- [7] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *INFOCOM, 2013 Proceedings IEEE*, April 2013, pp. 998–1006.
- [8] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 131–139, February 2014.
- [9] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *Communications, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, Aug. 2014.
- [10] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *Information Theory, IEEE Transactions on*, vol. 59, no. 12, pp. 8402–8413, Dec 2013.

- [11] J. Monserrat, H. Droste, O. Bulakci, J. Eichinger, O. Queseth, M. Stamatelatos, H. Tullberg, V. Venkatkumar, G. Zimmermann, U. Dotsch, and A. Osseiran, "Rethinking the mobile and wireless network architecture: The METIS research into 5G," in *Networks and Communications (EuCNC), 2014 European Conference on*, June 2014, pp. 1–5.
- [12] T. Han and N. Ansari, "On greening cellular networks via multicell cooperation," *IEEE Wireless Communications Magazine*, vol. 20, no. 1, pp. 82–89, 2013.
- [13] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 4, pp. 524–540, 2011.
- [14] T. Han and N. Ansari, "Powering mobile networks with green energy," *Wireless Communications, IEEE*, vol. 21, no. 1, pp. 90–96, February 2014.
- [15] Ericson Inc., "Sustainable energy use in mobile communications," Aug. 2007, white Paper.
- [16] T. Han and N. Ansari, "A traffic load balancing framework for software-defined radio access networks powered by hybrid energy sources," *CoRR*, vol. abs/1407.7780, 2014. [Online]. Available: <http://arxiv.org/abs/1407.7780>
- [17] —, "On optimizing green energy utilization for cellular networks with hybrid energy supplies," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 3872–3882, Aug. 2013.
- [18] L. Wang and G.-S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks: A tutorial," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 1, pp. 271–292, First 2013.
- [19] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 10–21, June 2011.
- [20] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 5, pp. 2484–2497, May 2013.
- [21] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *the 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2014)*, May 2014, pp. 37–42.
- [22] J. Zhou, M. Li, L. Liu, X. She, and L. Chen, "Energy source aware target cell selection and coverage optimization for power saving in cellular networks," in *Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications*, Hangzhou, China, Dec. 2010.
- [23] T. Han and N. Ansari, "Green-energy aware and latency aware user associations in heterogeneous cellular networks," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM'13)*, Atlanta, GA, USA, Dec 2013, pp. 4946–4951.
- [24] H. Goma, G. Messier, C. Williamson, and R. Davies, "Estimating instantaneous cache hit ratio using markov chain analysis," *Networking, IEEE/ACM Transactions on*, vol. 21, no. 5, pp. 1472–1483, Oct 2013.
- [25] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 1, Mar 1999, pp. 126–134 vol.1.
- [26] P. Rodriguez, C. Spanner, and E. W. Biersack, "Analysis of web caching architectures: Hierarchical and distributed caching," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, pp. 404–418, Aug. 2001.
- [27] P. Jelenkovic and A. Radovanovic, "Asymptotic insensitivity of least-recently-used caching to statistical dependency," in *the Twenty-Second Annual Joint Conference of the IEEE Computer and Communications.*, vol. 1, March 2003, pp. 438–447 vol.1.

- [28] Y. Zhang, N. Ansari, M. Wu, and H. Yu, "On wide area network optimization," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 4, pp. 1090–1113, Fourth 2012.
- [29] L. Kleinrock, *Queueing Systems: Computer applications*. Wiley-Interscience, 1976, ISBN: 978-0471491118.
- [30] A. Farbod and T. D. Todd, "Resource Allocation and Outage Control for Solar-Powered WLAN Mesh Networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 8, pp. 960–970, Aug. 2007.
- [31] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *Wireless Communications, IEEE*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
- [32] "Evolution of land mobile radio (including personal) communications: COST 231." [Online]. Available: <http://www.awe-communications.com/Propagation/Urban/COST/>
- [33] "HIT photovoltaic module." [Online]. Available: <http://us.sanyo.com/dynamic/product/Downloads/Panasonic%20HIT%20220A%20Data%20Sheet.pdf>
- [34] C. Riordan and R. Hulstron, "What is an air mass 1.5 spectrum? [solar cell performance calculations]," in *Photovoltaic Specialists Conference, 1990., Conference Record of the Twenty First IEEE*, May 1990, pp. 1085–1088.